Advancing Vision and Language in GI Diagnosis: Florence2 for Question Answering and Stable Diffusion for Image Synthesis

Notebook for ImageCLEFmedical at CLEF 2025

Krishna Tewari[1,*,†], Sukomal Pal[1,†]

[1]Indian Institute of Technology(BHU) Varanasi, India

Abstract

Recent advances in medical AI have underscored the importance of Visual Question Answering (VQA) and medical image generation. VQA systems enable automated reasoning over medical images using natural language queries, enhancing clinical interpretability. Meanwhile, generative models synthesize realistic medical images from textual descriptions, supporting data augmentation, simulation training, and rare case generation; particularly valuable in low-resource domains. Though evaluated independently in this challenge, these tasks are inherently complementary. VQA can aid in semantically annotating synthetic images, while synthetic images can enrich datasets to improve VQA model training. Together, they pave the way for robust multimodal diagnostic systems. In the ImageCLEFmed-MEDVQA-GI 2025 challenge, we address both subtasks: (1) closed-domain VQA and (2) medical image generation in the gastrointestinal (GI) domain. For VQA, we fine-tuned Microsoft's Florence2 vision-language transformer on the Kvasir-VQA dataset, using a custom preprocessing pipeline to remove specular highlights and black borders. Evaluation on the test sets yielded BLEU scores of 0.24/0.22, ROUGE-L scores of 0.87/0.88, and METEOR scores of 0.48/0.49, demonstrating strong domain-specific performance. For image generation, we fine-tuned Stable Diffusion with LoRA using synthetic captions produced by the QwQ language model. The model generated high-resolution (768×768) GI images. Evaluation on the private test set achieved a fidelity score of 0.2739, agreement of 0.739, and diversity of 0.6481, indicating high-quality synthesis. Our approach integrates fine-tuned VQA and diffusion models in a reproducible multimodal framework, advancing clinical image interpretation and dataset enrichment in low-resource GI healthcare.

Keywords

Medical VQA, ImageCLEFmed 2025, Multimodal AI, Clinical Question Answering, Synthetic GI Images, Specular Highlight Removal, Diffusion Models, Florence2, Low-Rank Adaptation (LoRA)

1. Introduction

In recent years, gastrointestinal (GI) image analysis has emerged as a cornerstone of diagnostic medicine due to the rising global burden of GI diseases. Colorectal cancer alone accounts for more than 1.9 million new cases and over 935,000 deaths annually, making it the third most deadly cancer worldwide[1]. High-resolution endoscopic techniques such as colonoscopy and capsule endoscopy generate vast quantities of visual data that require precise, real-time interpretation. Manual analysis is time-consuming and susceptible to variability. This motivates the need for automated, intelligent image interpretation systems. With the advancement of computational imaging, deep learning, and image-guided diagnostics, AI-driven tools now aid in the detection of polyps, ulcers, and inflammatory markers with increasing accuracy [1]. Additionally, the shift toward data-driven healthcare has accelerated the adoption of GI image analysis platforms which aim to reduce diagnostic delays and enhance early disease detection [2].

Visual Question Answering (VQA) plays a critical role in medical AI by bridging the gap between complex visual data and clinical decision-making through natural language interaction. In the medical

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

†These authors contributed equally.

krishnatewari.rs.cse24@itbhu.ac.in (K. Tewari); spal.cse@itbhu.ac.in (S. Pal)

0009-0005-6599-9956 (K. Tewari); 0000-0001-8743-9830 (S. Pal)

[1]https://www.who.int/news-room/fact-sheets/detail/cancer

domain, VQA systems are trained to interpret an image I and answer a clinician's question Q by producing an answer A, formalized as $f : (I, Q) \rightarrow A$. By leveraging state-of-the-art transformer architectures, these systems encode joint embeddings $\mathbf{z} = \phi(I, Q)$ that capture intricate relationships between image features and textual queries [3]. This facilitates interpretable and context-aware analysis of subtle pathological findings, which may be overlooked or misinterpreted during manual review. Consequently, VQA enhances diagnostic accuracy and efficiency, enabling clinicians to obtain rapid, and reliable answers that assist real-time decision-making.

In contrast, synthetic image generation addresses a complementary yet equally vital challenge: the scarcity and imbalance of annotated medical imaging datasets. Diffusion-based generative models learn to approximate the conditional distribution $p_\theta(\mathbf{x}_0|c)$ of realistic images $\mathbf{x}_0$ given clinical captions c by progressively denoising latent variables $\mathbf{z}_t$ over discrete time steps t [4]. This enables the creation of diverse, high-fidelity GI images that represent rare pathologies or underrepresented anatomical variations. Therefore, this circumvents the privacy issues inherent in using real data. These synthetic datasets not only expand the training resources for AI models but also allow systematic evaluation of diagnostic algorithms across a broader clinical spectrum. Importantly, when integrated with VQA pipelines, synthetic generation supports the creation of well-annotated multimodal data, reinforcing model robustness and clinical relevance [5].

Together, VQA and synthetic image generation serve distinct but synergistic functions within medical AI: VQA enhances clinical interpretability and decision support by enabling interactive image understanding, while synthetic generation expands and balances training data to improve model generalization and reliability.

To foster progress at this intersection of vision-language understanding and image synthesis, the ImageCLEFmed-MEDVQA-GI 2025 challenge [6] introduced a unique dual-task benchmark in the GI domain. This year's challenge focused on two core subtasks:

- **Subtask 1 (Closed-domain VQA):** Participants were required to develop models capable of answering medical questions grounded in GI endoscopic images, using datasets annotated with question-answer pairs [7]. It presented the need to leverage multimodal learning approaches that jointly analyze visual data and textual queries in order to provide responses.
- **Subtask 2 (Synthetic Image Generation):** This task involved generating synthetic GI images conditioned on clinical text prompts, such as descriptions of polyp types, anatomical landmarks, or procedural findings [5].

Both subtasks aim to address pressing challenges in the clinical AI landscape: Subtask 1 enhances interpretability and decision support, while Subtask 2 enables scalable data augmentation and training for low-resource tasks.

In response to this challenge, we participated in both subtasks independently by employing state-of-the-art models tailored to each task. For Subtask 1, we fine-tuned Microsoft's Florence2, a robust vision-language transformer, on the Kvasir-VQA dataset. To improve the quality of the visual input, we applied a specialized preprocessing pipeline addressing common endoscopic image artifacts such as specular highlights and black masks. Mixed-precision training was used to balance model performance and computational efficiency [8]. For Subtask 2, we fine-tuned Stable Diffusion v2-1 using Low-Rank Adaptation (LoRA), which allows efficient adaptation of large diffusion models with minimal resource demands [9]. This approach enabled the synthesis of high-resolution, clinically relevant GI images, enhancing the diversity and realism of generated medical data.

Through these efforts, we showcased the effectiveness of leveraging advanced vision-language and generative models for distinct yet complementary challenges in GI imaging, contributing to improved interpretability and data augmentation in clinical applications.

The rest of the paper is structured as follows. Section 2 provides a concise overview of prior research in this field. In Section 3, we provide the task overview and then describe the datasets used. Section 4 elaborates on our computational methodologies, and model specifications. Next, we explain the evaluation methodology, present our results and conduct a comprehensive analysis in Section 5. Finally, we conclude in Section 6.

2. Related Work

Recent advances in medical artificial intelligence have increasingly emphasized the integration of VQA and synthetic image generation to improve diagnostic accuracy, especially within the GI domain. This integration involves complex multimodal reasoning, combining visual data from medical images with natural language questions to generate clinically relevant answers.

Pioneering datasets like VQA-RAD [10] established early benchmarks by pairing radiological images with clinically meaningful questions and answers. This dataset enabled models to learn how to interpret medical images while simultaneously reasoning about associated clinical queries, focusing on radiology-specific pathologies. Building upon this, PathVQA [11] extended the framework to pathology by compiling over 32,000 question-answer pairs based on microscopic histopathological images. This allowed for deeper reasoning about cellular and tissue-level abnormalities, enhancing model capabilities to understand fine-grained pathological features. In the specific context of GI endoscopy, Kvasir-VQA [12] curated a large-scale dataset featuring endoscopic images paired with multiple question types, including anatomical, pathological, and procedural questions.

Simultaneously, the generation of synthetic medical images has become a critical technique for overcoming data scarcity and imbalance, particularly in rare disease classes that lack sufficient real-world examples. Generative Adversarial Networks (GANs) [13] represent a landmark approach wherein a generator network produces synthetic images, which are then judged by a discriminator network that learns to distinguish real from fake images. While GANs have been effective in many domains, their training often suffers from instability and mode collapse. This limits their capacity to produce diverse, high-fidelity medical images reliably.

To address these limitations, diffusion models [4, 14] have emerged as a powerful alternative. These models learn a reverse denoising process that transforms random noise into realistic images through a series of iterative refinement steps. Recent studies in medical imaging [15] demonstrate that diffusion models generate synthetic images with greater anatomical coherence and visual diversity than GANs, making them particularly suitable for augmenting GI imaging datasets.

On the front of vision-language integration, models like CLIP [16] learn joint embeddings of images and text by training on large-scale web data, enabling zero-shot image recognition and open-domain image-text reasoning. Similarly, BLIP (Bootstrapping Language-Image Pre-training) improves captioning and VQA by aligning vision and language features through a combination of contrastive and generative objectives. The Flamingo [17] model further extends this by enabling few-shot learning and open-ended multimodal reasoning using a large-scale transformer architecture.

In medical AI, domain-specific adaptations such as Med-Flamingo [18] have been proposed. Med-Flamingo fine-tunes large pretrained vision-language models on medical image-text pairs to generate clinical rationales. It also performs VQA with limited labeled data, capturing domain-specific semantic nuances. Despite these advances, these models often require extensive fine-tuning, which can be computationally prohibitive, and they may suffer from domain gaps because the base models are primarily trained on natural images and general language corpora.

To mitigate these challenges, parameter-efficient fine-tuning (PEFT) techniques such as LoRA [19] enable the adaptation of large pretrained models by training only a small number of additional low-rank parameter matrices. This significantly reduces the number of trainable parameters and computational resources while maintaining performance. PEFT approaches have shown promising results in medical imaging tasks [20, 15], allowing resource-efficient transfer learning but remain underexplored in complex multimodal GI VQA systems.

The recent ImageCLEFmed-MEDVQA challenges in 2023 [21] and 2024 [22] have driven state-of-the-art progress by encouraging innovative approaches to multimodal learning on GI endoscopic images and text queries. The 2023 winning team, UIT-Saviors [23], enhanced model performance by applying image enhancement techniques such as contrast adjustment and noise reduction to improve the visibility of fine endoscopic features before multimodal fusion with text. This preprocessing step enabled more precise visual feature extraction crucial for answering diagnostic questions. In 2024, the top solutions incorporated fine-tuned diffusion models for synthetic image augmentation alongside PEFT strategies

for adapting large multimodal transformers.

Collectively, these research efforts exemplify a converging paradigm of leveraging advanced multimodal architectures, computationally efficient fine-tuning, and high-fidelity synthetic data generation.

Building upon these foundations, our work addresses two seperate subtasks i.e. VQA and high-fidelity synthetic image generation within the GI tract. For VQA, we utilize the Florence generative vision-language model alongside a robust preprocessing pipeline tailored to endoscopic images. For image generation, we fine-tune diffusion models with LoRA for efficient adaptation. This dual-subtask framework enriches interpretability and data diversity, advancing the state-of-the-art in multimodal learning for GI healthcare applications.

3. Task Overview and Dataset

This section outlines the overall objectives of the task and presents a detailed description of the dataset used to perform our experimental analysis.

3.1. Task Overview

We participated in both Subtask 1 and Subtask 2 of the ImageCLEFmed-MEDVQA-GI 2025 challenge [24], which comprises the following components:

- **VQA:** This subtask requires generating textual answers based on image-question pairs. Participants must develop systems capable of interpreting endoscopic images in conjunction with corresponding clinical questions to produce accurate textual responses. For example, given an image showing a colon polyp and the question, *"Where in the image is the polyp located?"*, the system should output a relevant answer such as *"upper-left"* or *"in the center of the image."*
- **Image Generation:** This subtask involves building models that transform clinical text descriptions into high-fidelity synthetic GI images. The objective is to generate synthetic outputs that closely resemble real endoscopic images, such as those obtained through colonoscopy or gastroscopy, while preserving anatomical accuracy and clinical variability.

3.2. Dataset Information

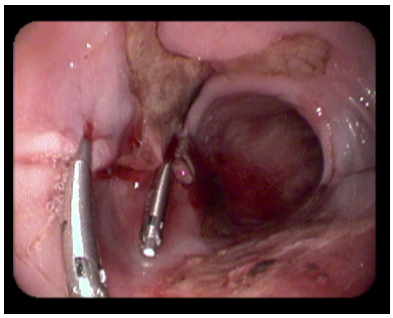
The dataset utilized in this challenge is *Kvasir-VQA* [12], a publicly available, expert-annotated dataset designed to advance research in medical VQA and related tasks within the GI diagnostic domain. The dataset is an extension of the *HyperKvasir* [25] and *Kvasir-Instrument* [26] datasets, enriched with natural language question-and-answer (QA) annotations to support multimodal learning tasks.

Kvasir-VQA consists of 6,500 high-resolution endoscopic images, each annotated with multiple clinically relevant questions spanning various anatomical regions and pathological findings of the GI tract. The dataset includes images of conditions such as *polyps, ulcers, esophagitis*, and scenes with *surgical instruments*, ensuring rich diversity for robust model training.

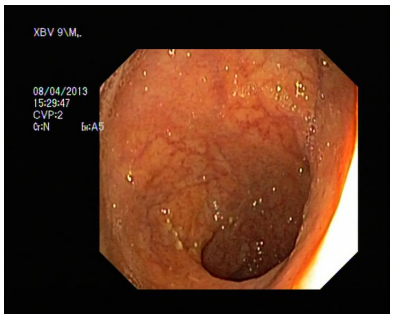
Each image is paired with multiple QA instances, resulting in over 38,500 VQA pairs. These questions cover:

- **Yes/No** queries (e.g., "Is there a polyp?"),
- **Categorical choices** (e.g., "What type of lesion is shown?"),
- **Spatial location** (e.g., "Where is the abnormality?"),
- **Numerical counts** (e.g., "How many instruments are visible?").

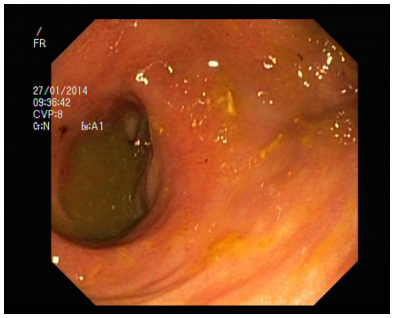
These QA pairs are grounded in real clinical reasoning, making the dataset well-suited for medical AI applications. The answer formats include *binary responses, categorical labels, ordinal values*, and *spatial descriptors*, simulating real-world diagnostic ambiguity and diversity. Figure 1 shows some sample images along with QA pairs.



Question: Where in the image is the instrument?
Answer: center; center-left; lower-center; lower-left



Question: Are there any abnormalities in the image? Check all that are present.
Answer: ulcerative colitis



Question: How many findings are present?
Answer: 1

Figure 1: Sample images with their associated QA pairs from the Development dataset

For the image generation subtask, Kvasir-VQA dataset is utilized, incorporating the same endoscopic images along with synthetically generated clinical captions[2]. Figure 2 shows a sample image and Table 1 shows sample synthetically generated captions associated to it. These captions simulate realistic diagnostic language and are designed to support conditional image synthesis tasks in the medical domain.

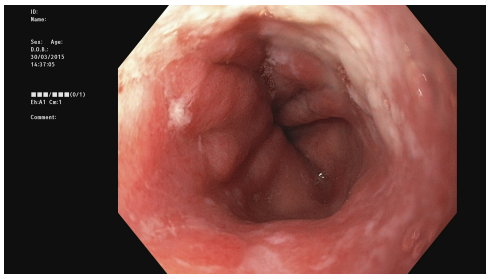


Figure 2: Example image from the Kvasir-VQA development dataset, showing an endoscopic view.

Table 1
Example of synthetically generated captions corresponding to the image in Figure 2 used as development for image generation subtask.

#	Synthetic Caption
1	A gastroscopy image showing esophagitis with no polyps present.
2	Two findings of esophagitis visible in the center and lower-center of the gastroscopy image.
3	Gastroscopy reveals oesophagitis with a z-line landmark in the center and lower-center.
4	The gastroscopy image displays esophagitis with pink, red, and white colorations; detection is challenging.
5	Esophagitis is evident in a gastroscopy image, marked by inflammation in multiple areas including center, center-left, and center-right.

[2]https://raw.githubusercontent.com/simula/ImageCLEFmed-MEDVQA-GI-2025/refs/heads/main/kvasir-captions.json

4. Methodology

This section presents our end-to-end methodology for solving the two core tasks: VQA and Image Generation. Inspired by recent advances in multimodal deep learning and guided by domain-specific requirements of GI imaging, we propose a system that combines vision-language fusion and generative modeling, supported by extensive preprocessing and PEFT.

4.1. VQA Pipeline (Subtask 1)

Our approach to the VQA task builds upon the Florence-2-base-ft model, a vision-language transformer designed for causal language modeling with integrated image and text understanding as shown in Figure 3. Each input pair consists of a clinically relevant natural language question and an associated colonoscopy or gastroscopy image. Prior to ingestion into the model, we perform crucial image preprocessing to address challenges inherent to endoscopic imaging.

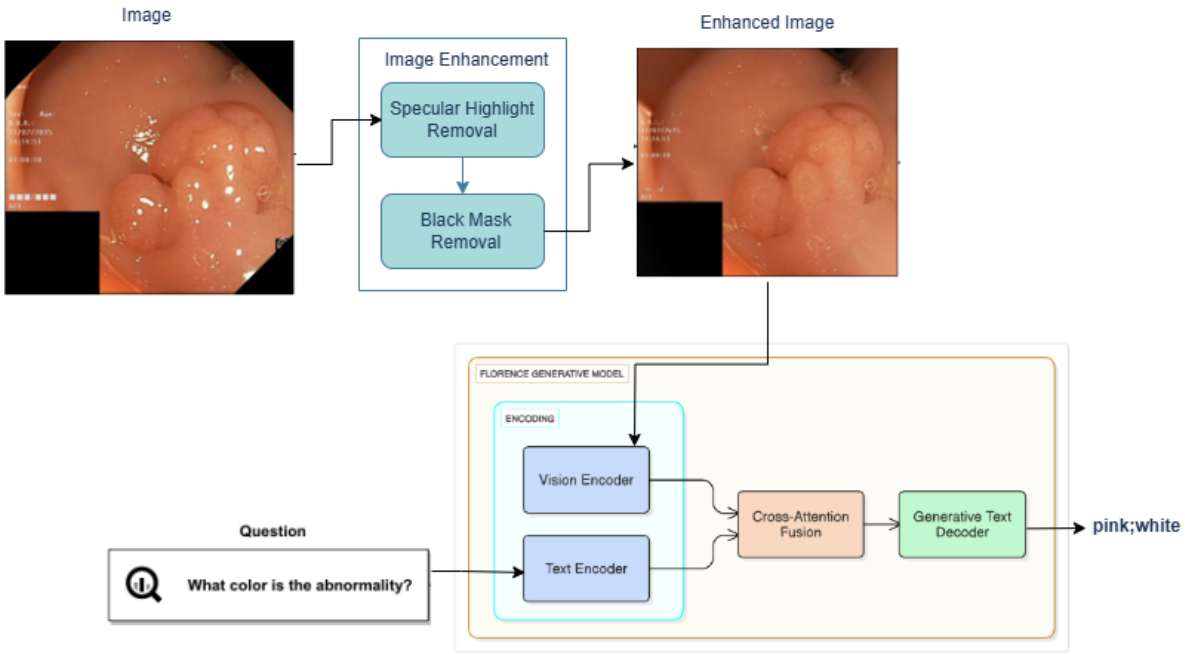


Figure 3: Step-by-step overview of the VQA pipeline utilizing Florence2 generative model.

4.1.1. Preprocessing

Colonoscopy images are often affected by optical and environmental artifacts that hinder automated image analysis. Two of the most common and disruptive issues include specular highlights, which are high-intensity reflections caused by wet mucosal surfaces illuminated by the endoscope's light source, and black masks, which are peripheral dark regions caused by circular lens vignetting or camera constraints. Both artifacts distort texture and shape cues vital for downstream tasks. Therefore, we developed a dedicated image enhancement module that performs targeted correction of these issues.

Specular Highlight Removal: Specular highlights are regions of over-saturation where the reflected light exceeds the sensor's dynamic range. These areas lack semantic information and often introduce noise into feature maps. We define the grayscale intensity of a given RGB image $I(x, y) \in \mathbb{R}^3$ as:

$$I_{\text{gray}}(x, y) = 0.299 \cdot R(x, y) + 0.587 \cdot G(x, y) + 0.114 \cdot B(x, y) \tag{1}$$

A binary mask M_s is computed by thresholding pixels above an empirically chosen intensity value $T_s = 240$:

$$M_s(x, y) = \begin{cases} 1 & \text{if } I_{\text{gray}}(x, y) \geq T_s \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

This mask is then dilated using a morphological operator to expand bright regions and fill small gaps:

$$M_d = M_s \oplus K \tag{3}$$

where K is a 5×5 square structuring element and $\oplus$ denotes dilation. We extract connected components and discard those with area $A < 100$ pixels to eliminate noise. The resulting regions are filled using Telea's inpainting algorithm [27], which solves the following Laplace equation with Dirichlet boundary conditions:

$$\text{div}(\nabla I^*(x, y)) = 0, \quad \forall (x, y) \in \Omega \tag{4}$$

where Ω is the inpainting domain and $I^*(x, y)$ is the reconstructed intensity map.

Black Mask Removal: Black masks occur due to the mismatch between the circular field of view and the rectangular image sensor. To detect these regions, we apply a low-threshold operation on I_{gray}:

$$M_b(x, y) = \begin{cases} 1 & \text{if } I_{\text{gray}}(x, y) \leq T_b \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

where $T_b = 5$ is the lower bound for black region detection. The bounding box of the foreground (non-black) region is computed as $(x_{\min}, x_{\max}, y_{\min}, y_{\max})$, and a padding margin $\delta = 5$ is added to avoid overcropping:

$$x'_{\min} = \max(0, x_{\min} - \delta), \quad x'_{\max} = \min(W, x_{\max} + \delta) \tag{6}$$

$$y'_{\min} = \max(0, y_{\min} - \delta), \quad y'_{\max} = \min(H, y_{\max} + \delta) \tag{7}$$

The resulting black regions are also reconstructed using Telea's method, ensuring a smooth transition between content and background.

These preprocessing methods significantly improved downstream VQA by normalizing image structure and reducing noise, consistent with findings in prior work [23, 28]. The threshold values T_s and T_b were selected through visual inspection over a random subset of development images to ensure maximal removal of unwanted artifacts while preserving diagnostically relevant features. Further automated tuning could be explored in future iterations. While additional enhancements such as contrast-limited adaptive histogram equalization (CLAHE), denoising, and sharpening were implemented, they were excluded from the final training phase to avoid overfitting to synthetic enhancements.

4.1.2. Model Architecture

Following preprocessing, images and tokenized text inputs are passed into the Florence-2[3] model. The image is processed through a transformer-based vision encoder, which extracts high-dimensional spatial features. Simultaneously, the question is encoded through a causal text encoder adapted for token-level language modeling. The two streams are fused via cross-attention layers within the model's architecture, allowing text tokens to dynamically attend to semantically relevant image regions. This fusion mechanism enables the model to align language components such as anatomical references, colors, or object types with visual counterparts in the image. Unlike traditional classification-based MedVQA systems, our model produces free-form text responses in a generative fashion. This is

[3] https://huggingface.co/microsoft/Florence-2-base

particularly advantageous for variable-length answers, such as anatomical descriptions or answers involving compound attributes (e.g., “Red and pink in the lower-left quadrant").

Training of the VQA model was performed using Hugging Face’s Trainer class. The model was fine-tuned over 3 epochs with a learning rate of 3e-5, using the AdamW optimizer with a weight decay factor of 0.01. Due to memory constraints and model size, we set the per-device batch size to 3 and used gradient accumulation over 16 steps. Mixed precision training (fp16) was enabled to optimize GPU memory usage. All training experiments were conducted on one NVIDIA H100 GPU (80GB VRAM), ensuring the model had sufficient computational capacity for large-scale multimodal learning. The final model was saved and uploaded to the Hugging Face Model Hub under the repository krissTewari/Florence-2-vqa-final[4].

4.2. Image Generation Pipeline (Subtask 2)

For Subtask 2, we developed a dedicated image generation pipeline capable of synthesizing realistic endoscopic images from structured diagnostic captions as shown in Figure 4.

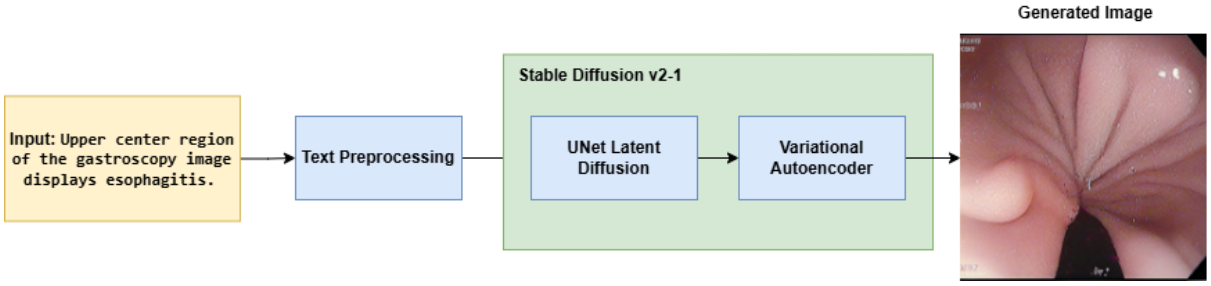


Figure 4: Step-by-step overview of the image generation pipeline utilizing Stable Diffusion v2-1.

We used Stable Diffusion v2-1[5] as the base generative model, a state-of-the-art latent diffusion model pretrained on billions of text-image pairs. Unlike traditional pixel-space generators, Stable Diffusion operates within a latent space learned by a variational autoencoder (VAE), which enables high-resolution generation with reduced computational overhead. The model was configured to generate images at a resolution of 768×768 pixels, which balances visual fidelity and computational feasibility in the medical domain, especially where fine anatomical structures must be preserved.

The dataset used for fine-tuning consisted of caption-image pairs from the development dataset of the ImageCLEFmed-MEDVQA-GI task. These captions are natural language descriptions of clinically relevant image content, including the presence, size, shape, color, and location of abnormalities (e.g., polyps), as well as the presence of instruments or anatomical landmarks.

To assess the best-suited backbone for domain-specific image synthesis, we initially evaluated three model variants: Stable Diffusion v1-5[6], Stable Diffusion XL (SDXL)[7], and Stable Diffusion v2-1. While SDXL demonstrated promising compositional quality due to its expanded architecture and 1024×1024 latent space, it incurred significantly higher VRAM requirements and longer training times, making it impractical for constrained medical datasets. Stable Diffusion v1-5, while computationally efficient, showed limitations in accurately capturing fine-grained medical textures and anatomical features. In contrast, Stable Diffusion v2-1 offered a strong trade-off between resolution (768×768), performance, and memory usage, and thus was selected for the final training and submission phase.

In order to adapt the pretrained Stable Diffusion model to this highly specialized domain while avoiding the need to update the full set of model parameters, we employed LoRA. During training, only the inserted matrices are updated, while the base model weights remain frozen. In our implementation,

[4]https://huggingface.co/krissTewari/Florence-2-vqa-final
[5]https://huggingface.co/stabilityai/stable-diffusion-2-1
[6]https://huggingface.co/runwayml/stable-diffusion-v1-5
[7]https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0

LoRA was configured with a rank of 4. This choice was guided by prior work and common practice in parameter-efficient fine-tuning (PEFT) for diffusion models. We also conducted preliminary experiments with LoRA ranks of 2, 4, and 8. Rank 4 provided the best trade-off between parameter efficiency and generation quality without exceeding memory constraints. Higher ranks showed minimal improvement in fidelity but significantly increased VRAM usage. We applied LoRA modules specifically to the cross-attention layers within the UNet, which allows the model to effectively learn how to condition the image synthesis on medical language prompts without catastrophic forgetting of general-domain image priors. This method significantly reduced VRAM usage and made training feasible on commercially available hardware without degrading generative quality.

The training process was orchestrated using Hugging Face's accelerate launcher in conjunction with the diffusers and peft libraries. Training was conducted over 3 full epochs, with a batch size of 4 and gradient accumulation over 4 steps. We used a constant learning rate of 1e-4, which was selected based on prior literature and empirical performance on LoRA tuning. The training loop included automatic checkpointing every 500 steps, with a maximum of three recent checkpoints retained to support resumption and rollback. To monitor the qualitative progression of generation, we used validation prompts during training. These prompts were constructed to mirror common VQA queries, one of which is: *"The colonoscopy image contains a single, moderate-sized polyp that has not been removed, appearing in red and pink tones in the center and lower areas.* This prompt captures both anatomical features and visual attributes and serves as a reliable baseline for visual inspection during training.

All training was carried out on a server equipped with four NVIDIA L40s GPUs, each offering 48 GB of VRAM. The model was trained using mixed-precision (fp16) to reduce memory consumption and improve throughput. The use of gradient checkpointing within the UNet further minimized the peak memory footprint. Upon completion of training, the final LoRA adapter weights were pushed to the Hugging Face Model Hub under the repository krissTewari/sd-kvasir-imagen-demo[8], making them publicly available for reproducibility and further research use.

5. Results

In this section, we present a detailed evaluation of our proposed models on VQA and Image Generation tasks. The assessment is conducted using standard evaluation metrics across both public and private test sets to provide a comprehensive understanding of model performance.

5.1. Evaluation Metrics

The performance of our models was assessed using a comprehensive set of evaluation metrics tailored to the distinct characteristics of VQA and Image Generation tasks. These metrics were selected to provide a multidimensional understanding of linguistic quality, semantic alignment, and visual fidelity.

5.1.1. Metrics for VQA

- **BLEU (Bilingual Evaluation Understudy):** BLEU [29] measures the modified n-gram precision of generated answers with respect to reference answers, incorporating a brevity penalty to discourage short outputs:

$$\text{BLEU-N} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{8}$$

where

$$BP = \begin{cases} 1 & \text{if } c > r \\ e^{(1-r/c)} & \text{if } c \leq r \end{cases} \tag{9}$$

[8] https://huggingface.co/krissTewari/sd-kvasir-imagen-demo

c is the length of the candidate, r is the reference length, p_n is the modified n-gram precision, and w_n are weights (typically uniform).

- **ROUGE (Recall-Oriented Understudy for Gisting Evaluation):** ROUGE [30] is a family of metrics that evaluate n-gram and sequence overlaps. We employ ROUGE-1, ROUGE-2, and ROUGE-L, where:

$$\text{ROUGE-N} = \frac{\sum_{\text{gram}_n \in \{\text{Ref}\}} \min(\text{Count}_{\text{match}}(\text{gram}_n), \text{Count}_{\text{cand}}(\text{gram}_n))}{\sum_{\text{gram}_n \in \{\text{Ref}\}} \text{Count}_{\text{Ref}}(\text{gram}_n)} \tag{10}$$

  and ROUGE-L measures the Longest Common Subsequence (LCS) recall.
- **METEOR (Metric for Evaluation of Translation with Explicit ORdering):** METEOR [31] combines unigram precision and recall with stemming, synonyms, and a fragmentation penalty:

$$\text{METEOR} = F_{\text{mean}} \cdot (1 - P_{\text{frag}}) \tag{11}$$

  where F_{mean} is the harmonic mean of precision and recall, and P_{frag} penalizes fragmented matches.

5.1.2. Metrics for Image Generation

5.1.3. Metrics for Image Generation

In addition to expert ratings, four automated quantitative metrics are used to assess the realism, consistency, and variability of generated images. These metrics leverage BiomedCLIP [**?**] image embeddings, providing a semantically grounded evaluation tailored to the medical imaging domain.

- **Fidelity (↑):** Measures how realistic the generated images are compared to real colonoscopy images. It is computed as a scaled inverse of the Fréchet Inception Distance (FID), using BiomedCLIP embeddings:

$$\text{Fidelity} = \frac{1000}{1 + \text{mean-FID}(A_i, R_i)} \tag{12}$$

  where A_i and R_i denote the generated and real image features for prompt i. Higher scores indicate greater similarity to real images.
- **Agreement (↑):** Evaluates semantic and visual consistency between images generated from original and rephrased prompts. Agreement is calculated as the average cosine similarity between BiomedCLIP embeddings of images from paired prompts:

$$\text{Agreement} = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{|A_i||B_i|} \sum_{a \in A_i, b \in B_i} \frac{a \cdot b}{\|a\|\|b\|} \tag{13}$$

  where A_i and B_i represent the image embedding sets from the original and rephrased prompts.
- **Diversity (↑):** Quantifies intra-prompt variability among generated images, promoting diverse outputs rather than mode collapse. For each prompt, we compute the average pairwise Euclidean distance between the BiomedCLIP embeddings of the 10 generated images:

$$\text{Diversity} = \frac{1}{M} \sum_{i=1}^{M} \text{pdist}(F_i) \tag{14}$$

  where F_i is the set of normalized embeddings per prompt and pdist denotes the mean pairwise distance function.
- **Fréchet BiomedCLIP Distance (FBD) (↓):** Assesses global distributional similarity between the entire set of generated and real images using the Fréchet distance, computed in the BiomedCLIP embedding space:

$$\text{FBD} = \|\mu_{\text{gen}} - \mu_{\text{real}}\|^2 + \text{Tr}\left(\Sigma_{\text{gen}} + \Sigma_{\text{real}} - 2(\Sigma_{\text{gen}}\Sigma_{\text{real}})^{1/2}\right) \tag{15}$$

  where μ and Σ are the mean and covariance of embeddings from generated and real images, and Tr denotes the matrix trace. Lower scores indicate better global alignment.

5.2. Results

This section presents the evaluation outcomes and detailed analysis of our approaches.

5.2.1. VQA

Table 2 presents the performance of our Florence2 model on the public and private test sets.

Table 2
Florence2 Model Performance on Public and Private Test Sets

Set	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	METEOR
Public	0.24	0.87	0.11	0.87	0.48
Private	0.22	0.88	0.11	0.88	0.49

As shown, the model achieves strong and consistent results across both datasets. On the public test set, it attains a BLEU score of 0.24, ROUGE-1 and ROUGE-L scores of 0.87, while ROUGE-2 and METEOR scores of 0.11 and 0.48, respectively. Performance on the private test set is comparably robust, with slight variations showing BLEU at 0.22 and marginal improvements in ROUGE-1, ROUGE-L, and METEOR scores. These results confirm the model's reliable generalization and effectiveness in answering domain-specific medical questions in GI imagery.

Table 3 reports our ablation study i.e., the performance of the Florence2 model under different configurations involving LoRA fine-tuning and image enhancement.

When no LoRA fine-tuning or image enhancement is applied, the model achieves a BLEU score of 0.20, METEOR of 0.24, and ROUGE-L and ROUGE-1 scores near 0.45 and 0.46, respectively. Notably, applying LoRA fine-tuning without prior image enhancement led to a significant drop in performance. This suggests that the Florence2 model, which is pre-trained on clean natural images, may overfit to visual noise when exposed to unprocessed endoscopic artifacts. The model might mistakenly associate specular highlights or dark borders with specific answers, introducing spurious correlations. Therefore, preprocessing appears essential before domain-specific fine-tuning, serving as a form of regularization.

Applying image enhancement without LoRA fine-tuning improves results moderately, with BLEU rising to 0.15 and ROUGE-L to 0.28, demonstrating the positive effect of improved visual inputs on answer generation quality. The best results occur when LoRA fine-tuning is combined with image enhancement, achieving a BLEU score of 0.24, METEOR of 0.48, and ROUGE-L and ROUGE-1 scores of 0.87 each, illustrating strong synergy between model adaptation and high-quality input preprocessing.

In the ablation study, "With Image Enh." refers exclusively to the application of specular highlight and black mask removal. Other image enhancement techniques like CLAHE, sharpening, and denoising were initially explored but later excluded from the final configuration. This decision was based on qualitative assessments and pilot experiments, which showed negligible performance gains and a risk of overfitting to artificially enhanced features.

These findings emphasize that image enhancement consistently benefits VQA performance, while LoRA fine-tuning is most effective when integrated with such preprocessing, ultimately enhancing the domain-specific interpretability and accuracy of the Florence2 model on GI medical images.

Table 3
Florence2 Model Results with different settings

Setting	BLEU	METEOR	ROUGE-L	ROUGE-1	ROUGE-2
Without LoRA, No Image Enh.	0.20	0.24	0.45	0.46	0.09
With LoRA, No Image Enh.	0.10	0.09	0.13	0.13	0.01
Without LoRA, With Image Enh.	0.15	0.15	0.28	0.28	0.05
With LoRA, With Image Enh.	**0.24**	**0.48**	**0.87**	**0.87**	**0.11**

5.2.2. Image Generation

Table 4 summarizes the performance of our fine-tuned Stable Diffusion v2-1 across the public and private test sets.

Table 4
Performance of the Image Generation Model on Public and Private Test Sets

Set	Fidelity	Agreement	Diversity	FBD
Public	0.27	0.74	0.66	1923.16
Private	0.2739	0.739	0.6481	1694.97

The fidelity scores of approximately 0.27 on both datasets indicate that the generated images closely resemble real GI images in terms of visual quality and clinical relevance. Agreement scores near 0.74 demonstrate strong alignment between the generated images and the corresponding clinical captions, validating the model's ability to synthesize medically coherent images. Diversity scores above 0.64 reflect a healthy variety in the generated samples, important for capturing the range of possible clinical presentations. The FBD values, which measure the distributional similarity between generated and real image features, are lower on the private set (1694.97) compared to the public set (1923.16), suggesting better realism and fidelity in the private evaluation. Overall, these results highlight the effectiveness of our LoRA-fine-tuned Stable Diffusion approach in producing high-resolution, clinically meaningful GI images from textual descriptions.

To evaluate the influence of different base models on image generation performance, we conducted an ablation study comparing Stable Diffusion v1-5, SDXL, and Stable Diffusion v2-1 (our main model) on the public test set. The results in Table 5 demonstrate that Stable Diffusion v2-1 outperforms other variants across critical metrics.

Table 5
Different Models' Image Generation Performance on Public Test Set

Model	Fidelity	Agreement	Diversity	FBD
Stable Diffusion v1-5	0.26	0.69	0.74	1742.08
SDXL	0.28	0.68	0.65	1950.30
Stable Diffusion v2-1	**0.27**	**0.74**	**0.66**	**1923.16**

Stable Diffusion v2-1 achieves the highest agreement (0.74) score, indicating superior alignment with clinical captions, while maintaining competitive fidelity (0.27) and diversity (0.66). SDXL demonstrates a slightly better fidelity (0.28) but lower agreement and diversity, suggesting it produces high-quality images that are somewhat less relevant or varied. Stable Diffusion v1-5 shows higher diversity (0.74) but comparatively lower fidelity and agreement. The Fréchet Brain Distance (FBD) values indicate that v2-1 and v1-5 generate more realistic images compared to SDXL. Overall, these results support the conclusion that Stable Diffusion v2-1 is the best balanced model for GI medical image generation.

Visual Samples: Figure 5 showcases example outputs generated by our top-performing image generation methodology, illustrating the quality and clinical relevance of the synthetic GI images.

6. Conclusion and Future Work

This paper presents two distinct contributions: a VQA system and a synthetic image generation model, each addressing separate subtasks within the GI medical imaging domain.

For the VQA subtask, we fine-tuned the *Florence2* vision-language transformer on the Kvasir-VQA dataset. The model was trained to generate accurate textual answers to clinical questions posed over endoscopic images. To improve input quality, we introduced preprocessing techniques to remove specular highlights and black borders, common artifacts in endoscopic videos. This enhanced the

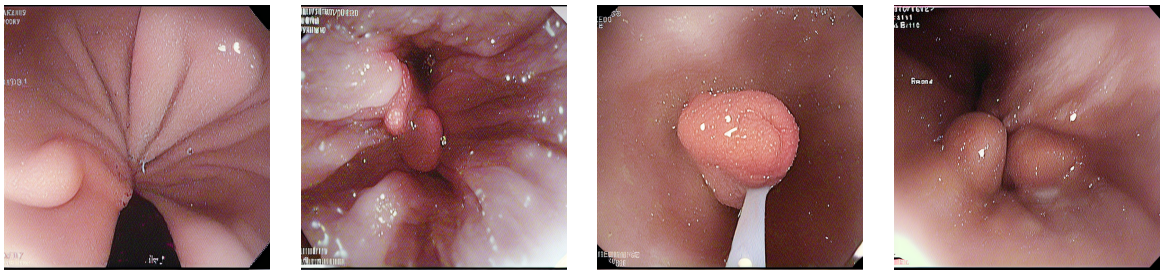


Figure 5: Examples of synthetic GI images generated by our diffusion-based model using clinical prompts.

model's ability to align visual and textual features, resulting in strong performance across standard language generation metrics including BLEU, ROUGE-L, and METEOR.

For the image generation subtask, we fine-tuned *Stable Diffusion v2-1* using LoRA, enabling the generation of high-resolution GI images from structured clinical captions. The model demonstrated a strong ability to produce visually diverse and semantically accurate outputs, with favorable scores across fidelity, agreement, and diversity metrics. This highlights the potential of diffusion models for synthetic data generation in low-resource medical settings.

Future Work. Although the two tasks are independent, their outcomes suggest several promising avenues for the research community, some of which are outlined below:

- **Multilingual VQA systems:** Expanding VQA models to handle clinical questions in multiple languages could increase accessibility and support global deployment of AI-assisted diagnostics.
- **Federated learning for privacy-preserving model training:** Training vision-language and generative models across distributed medical centers without transferring raw patient data would enable broader collaboration while adhering to privacy regulations.
- **Cross-task synergy:** Although VQA and image generation were addressed as separate subtasks, their interdependence presents a promising direction. For example, synthetic images generated from clinical captions could be validated or further annotated by VQA models, enabling the construction of weakly supervised multimodal datasets. Conversely, VQA outputs may guide or condition future image generation pipelines. Exploring such interactions could lead to more robust and context-aware medical AI systems.

Continued research in these directions can contribute to the development of scalable, trustworthy, and generalizable AI tools for GI diagnostics and other domains of medical imaging.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] M. A. Berbís, J. Aneiros-Fernández, F. J. Mendoza Olivares, E. Nava, A. Luna, Role of artificial intelligence in multidisciplinary imaging diagnosis of gastrointestinal diseases, World Journal of Gastroenterology 27 (2021) 4395–4412. URL: https://doi.org/10.3748/wjg.v27.i27.4395. doi:10.3748/wjg.v27.i27.4395.

[2] I. Barua, P. Vennapusa, O. Al-Jibury, V. Subramanian, V. R. Muthusamy, B. Hayee, Effectiveness of artificial intelligence-assisted colonoscopy in early diagnosis of colorectal cancer: A systematic

review, eClinicalMedicine 58 (2023) 101875. URL: https://doi.org/10.1016/j.eclinm.2023.101875. doi:10.1016/j.eclinm.2023.101875.
[3] L. Jing, X. Xie, A. Wong, Medvqa: Advances in medical visual question answering, IEEE Transactions on Medical Imaging 42 (2023) 1234–1248. doi:10.1109/TMI.2023.3245678.
[4] J. Ho, A. Jain, P. Abbeel, Denoising diffusion probabilistic models, in: NeurIPS, 2020.
[5] I. Ejiga, R. Smith, H. Nguyen, Synthetic medical image generation and visual question answering: A multimodal pipeline for clinical ai, Computers in Science and Engineering 36 (2024) 56–72. doi:10.1109/CompSciEng.2024.1023456.
[6] B. Ionescu, H. Müller, D.-C. Stanciu, A. Idrissi-Yaghir, A. Radzhabov, A. G. S. de Herrera, A. Andrei, A. Storås, A. B. Abacha, B. Bracke, B. Lecouteux, B. Stein, C. Macaire, C. M. Friedrich, C. S. Schmidt, D. Fabre, D. Schwab, D. Dimitrov, E. Esperança-Rodier, G. Constantin, H. Becker, H. Damm, H. Schäfer, I. Rodkin, I. Koychev, J. Kiesel, J. Rückert, J. Malvehy, L.-D. Ștefan, L. Bloch, M. Potthast, M. Heinrich, M. A. Riegler, M. Dogariu, N. Codella, P. H. P. Nakov, R. Brüngel, R. A. Novoa, R. J. Das, S. A. Hicks, S. Gautam, T. M. G. Pakull, V. Thambawita, V. Kovalev, W.-W. Yim, Z. Xie, Overview of imageclef 2025: Multimedia retrieval in medical, social media and content recommendation applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Proceedings of the 16th International Conference of the CLEF Association (CLEF 2025), Springer Lecture Notes in Computer Science LNCS, Madrid, Spain, 2025.
[7] B. Ionescu, H. Müller, A.-M. Drăgulinescu, J. Rückert, A. Ben Abacha, A. García Seco de Herrera, L. Bloch, R. Brüngel, A. Idrissi-Yaghir, H. Schäfer, Overview of the ImageCLEF 2024: Multimedia Retrieval in Medical Applications, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer, Cham, Switzerland, 2024, pp. 140–164. doi:10.1007/978-3-031-71908-0_7.
[8] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh, Mixed precision training, in: International Conference on Learning Representations (ICLR), 2018.
[9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, in: International Conference on Learning Representations (ICLR), 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.
[10] J. Lau, E. Lehman, P. Sharma, A dataset and exploration of models for understanding radiology reports, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2018, pp. 1–10. URL: https://aclanthology.org/D18-1001/.
[11] X. He, Y. Zhang, L. Mou, E. Xing, P. Xie, Pathvqa: 30000+ questions for medical visual question answering, arXiv preprint arXiv:2003.10286 (2020).
[12] S. Gautam, A. M. Storås, C. Midoglu, S. A. Hicks, V. Thambawita, P. Halvorsen, M. A. Riegler, Kvasir-VQA: A Text-Image Pair GI Tract Dataset, in: ACM Conferences, Association for Computing Machinery, New York, NY, USA, 2024, pp. 3–12. doi:10.1145/3689096.3689458.
[13] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems (NeurIPS) 27, NeurIPS, 2014, pp. 2672–2680.
[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: CVPR, 2022, pp. 10684–10695.
[15] C. Lian, H.-Y. Zhou, Y. Yu, L. Wang, Less could be better: Parameter-efficient fine-tuning advances medical vision foundation models, arXiv preprint arXiv:2401.12215 (2024).
[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, I. Sutskever, Learning transferable visual models from natural language supervision, in: Proceedings of the 38th International Conference on Machine Learning (ICML), PMLR, 2021, pp. 8748–8763. URL: https://proceedings.mlr.press/v139/radford21a.html.
[17] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, C. Reeves, A. Zisserman, K. Kavukcuoglu, C. Pal, K. Millican, D. Dohan, A. Mensch, S. Cabi, J. Menick, P.-S. Huang, B. Beyret, T. Le Paine, O. Vinyals, A. Brock, K. Simonyan, Flamingo: a visual language model for few-shot learning, arXiv preprint arXiv:2204.14198 (2022).

[18] M. Moor, Q. Huang, S. Wu, M. Yasunaga, Y. Dalmia, J. Leskovec, C. Zakka, E. P. Reis, P. Rajpurkar, Med-flamingo: A multimodal medical few-shot learner, in: Proceedings of the 3rd Machine Learning for Health Symposium, PMLR, 2023, pp. 353–367. URL: https://proceedings.mlr.press/v225/moor23a.html.

[19] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Chen, W. Chen, Lora: Low-rank adaptation of large language models, in: Proceedings of the 38th International Conference on Machine Learning (ICML), PMLR, 2021, pp. 11113–11125.

[20] R. Dutt, L. Ericsson, P. Sanchez, S. A. Tsaftaris, T. Hospedales, Parameter-efficient fine-tuning for medical image analysis: The missed opportunity, in: N. Burgos, C. Petitjean, M. Vakalopoulou, S. Christodoulidis, P. Coupe, H. Delingette, C. Lartizien, D. Mateus (Eds.), Proceedings of The 7th International Conference on Medical Imaging with Deep Learning (MIDL), volume 250 of *Proceedings of Machine Learning Research*, PMLR, 2024, pp. 406–425. URL: https://proceedings.mlr.press/v250/dutt24a.html.

[21] S. A. Hicks, A. Storås, P. Halvorsen, T. de Lange, M. A. Riegler, V. Thambawita, Overview of imageclefmedical 2023 – medical visual question answering for gastrointestinal tract, in: CLEF2023 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Thessaloniki, Greece, 2023.

[22] S. A. Hicks, A. Storås, P. Halvorsen, M. A. Riegler, V. Thambawita, Overview of imageclefmedical 2024 – medical visual question answering for gastrointestinal tract, in: CLEF2024 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Grenoble, France, 2024.

[23] T. M. Thai, A. T. Vo, H. K. Tieu, L. N. P. Bui, T. T. B. Nguyen, Uit-saviors at medvqa-gi 2023: Improving multimodal learning with image enhancement for gastrointestinal visual question answering, in: Proceedings of the MEDVQA-GI Workshop 2023, 2023. URL: https://arxiv.org/abs/2307.02783, arXiv preprint arXiv:2307.02783.

[24] S. Gautam, P. Halvorsen, M. A. Riegler, V. Thambawita, S. A. Hicks, Overview of imageclefmedical 2025 – medical visual question answering for gastrointestinal tract, in: CLEF2025 Working Notes, CEUR Workshop Proceedings, CEUR-WS.org, Madrid, Spain, 2025.

[25] H. Borgli, V. Thambawita, P. H. Smedsrud, S. Hicks, D. Jha, S. L. Eskeland, K. R. Randel, K. Pogorelov, M. Lux, D. T. D. Nguyen, D. Johansen, H. K. Johansen, M. A. Riegler, P. Halvorsen, HyperKvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy, Scientific Data 7 (2020) 1–14. doi:10.1038/s41597-020-00622-y.

[26] D. Jha, S. Ali, K. Emanuelsen, S. A. Hicks, V. Thambawita, E. Garcia-Ceja, M. A. Riegler, T. de Lange, P. T. Schmidt, D. Johansen, Kvasir-instrument: Diagnostic and therapeutic tool segmentation dataset in gastrointestinal endoscopy, in: MultiMedia Modeling, Springer, 2021, pp. 218–229. URL: https://datasets.simula.no/kvasir-instrument/. doi:10.1007/978-3-030-67835-7_18.

[27] A. Telea, An image inpainting technique based on the fast marching method, Journal of graphics tools 9 (2004) 23–34.

[28] Y. Kumar, B. Verma, R. Srivastava, Gastrointestinal abnormality detection in wireless capsule endoscopy images using deep learning, Computerized Medical Imaging and Graphics 79 (2020) 101678.

[29] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), 2002, pp. 311–318.

[30] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text Summarization Branches Out: Proceedings of the ACL-04 Workshop, 2004, pp. 74–81.

[31] S. Banerjee, A. Lavie, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments, in: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, 2005, pp. 65–72.